| Mitwirkende	
Børre Gaup	AIM: boerregaup E-Mail: borre.gaup@samediggi.no
Ciprian Gerstenberger	AIM: ciprian.gerstenberger@gmail.com E-Mail: ciprian.gerstenberger@uit.no
Maja Lisa Kappfjell	
Sjur	AIM: sjurnm@mac.com E-Mail: sjur.moshagen@samediggi.no
Thom	AIM: thomasomma
Tomi	AIM: topieski E-Mail: tomi.pieski@hum.uit.no
Trond Trosterud	AIM: ttrosterud E-Mail: trond.trosterud@hum.uit.no

!!!*Meeting setup*

* Date: 23.8.2010
* Time: 09.30 Norw. time
* Place: Internet
* Tools: SubEthaEdit, iChat

!!!*Agenda*

Cf. one of the following, depending on context:
* the upper bar of the SEE window (provided you use the JSPWiki syntax mode)
* the TOC in Forrest-rendered output, like HTML and PDF

!!!*Opening, agenda review, participants*

* Opened at 10:01.
  S.
* Present: __Børre, Ciprian, Maja, Sjur, Thomas, Tomi, Trond__
* Absent: __none__
* Agenda accepted as is

!!!*Updated task status since last meeting*

!!*Børre*
* get all the latest Sámi Parliament protocolls into our repository
** Found protocols in nob-sme versions up to 2005, have added most of them.
Børre Gaup
   Protocols from 2006 up to the second protocol in 2010 don't seem to have
Børre Gaup
   been translated, so they are lacking.
Børre Gaup
* contact the Sámi translators at the government and ministeries to get more
  parallell texts
** not done
Børre Gaup
* corpus access on the XServe
** not done
Børre Gaup
* turn on corpus summary crontab
** not done
Børre Gaup
* implement language switch for static divvun site
** not done
Børre Gaup
* improve XSL script to transform leaflet Forrest XDocs to an OOo Draw document
** not done
Børre Gaup
* get translations of thank-you letter
** not done
Børre Gaup
* make the new SL Server services functional:
** group calendars

*** not done
Børre Gaup

* set up corpus mirroring on the XServe again
** not done
Børre Gaup

* give Maja a WEbSak intro
** not done
Børre Gaup

* add checkout of the free corpus to the gtsetup.sh script
** not done
Børre Gaup

* add checkout of the bound corpus when on victorio
** not done
Børre Gaup

* {{sma}} seminar in August/September:
** test installation of all needed software
** test XMLEditor on Windows, including plugin installation
** test svn checkout and patch creation on Windows, using a GUI client
*** not done
Børre Gaup

* [fix bugs!|http://giellatekno.uit.no/bugzilla]
** Fixed #860
Børre Gaup

!!Ciprian
* make risten2 budget proposal to the SD with __Sjur__
** done
Ciprian …

* {{sma}} seminar in August/September:
** prepare slides
*** ongoing
Ciprian Gerste…

** test XMLEditor plugin installation
*** todo
Ciprian G…

** augment the dictionary DTD with oahpa features and requirements
*** waiting for Sjur's counterarguments (__Trond__ and __Lene__ agree on the intermediate solution proposed by __Cip__)
Ciprian Gerstenberger

** finnish the setup of the sma-oahpa on victorio
*** ongoing
Ciprian Gerste…

** build a preliminary data set for sma-Leksa (based on the DTD)
*** ongoing
Ciprian Gerste…

* Skolt-Oahpa:
** fine tuning of semantic tags
** finish localization
*** todo
Ciprian G…

* terminology:
** merged the 2 doc files with the xls file containing the 2006 law terminology
   db
** prepare a StarDict test version of the merged data for demo
** compile a dictionary in StarDict format with all data featuring the
   properties wished by Ellen Ravna
*** todo
Ciprian G…

* add license header to all files
* PhD-Projektbeskrivelse (internal deadline end of June 2010):
** search, read, write (execute): high priority!
* read and answer the newsgroups messages
*** reading Performance Grammar and Nunberg's Linguistics of Punctuation
Ciprian Gerstenberger

* update corpora (both monolingual and parallel) for Oslo (Glossa)
** check the correctness of parallelity between sme and nob files
*** after setting up the whole conversion pipeline
Ciprian Gerstenberger

** sentence-align sme with nob: fix tca2 problem
*** todo
Ciprian G…

** analyse/disambiguate: todo -- waiting for the last version of FSTs
* corpus infra:
** check the conversion from old rcs corpus repository to new svn-based
** make the conversion scripts work with the new svn repository
*** adjusted paths to scrips after reorganizing script-dir
Ciprian Gerstenberger
*** add check for almost empty content: todo
Cipria...
*** add processing of svg files: todo
Cipria...
** reorganise subdirs as needed
*** created a subdir corpus and moved all (check with __Børre__, __Trond__ and __Sjur__) relevant file there
Ciprian Gerstenberger
*** ajsusted paths in the analye file, now checking the convert script
Ciprian Gerstenberger
* infrastructure
** test cwb
** continue with restructuring and cleaning the script catalogue as suggested
   in the newsgroup
*** done for corpus-related stuff
Ciprian Gerstenberger
** transform sme-lexC files into XML format
*** apparently, there is a script that does exactly that (ask __Tomi__)
Ciprian Gerstenberger
** make a schema/dtd description of the lexC-file (experiment with
   Komi and Romanian)
*** todo
Ciprian G...
* GT web:
** add a tree visualizer for the dependency trees
** add input help for special characters on the tool sites
** automatise the web statistics
** filter (English, German, etc.) input using language detection tools
** put a note on the sites that these are NOT MT tools
** input help for generating wordforms (dropdown menus).
*** all here: todo
Ciprian Gerstenberger
* Sandbox Oahpa:
** integrate reCAPTCHA into Django (as __Lene__ suggested)
** debug the installed sb_oahpa
** update and correct the Oapha docu site
*** all here: todo
Ciprian Gerstenberger
* Running Oahpa:
** implement a testbench for Vasta (as __Lene__ needs)
** integrate reCAPTCHA after the SB-test
** add an Oahpa clock and date excercise (cf. Numra)
** email notification when the server goes down
** check the XXX?
*** all here: todo
Ciprian Gerstenberger
* Sjd/Kom/Sms/Etc-Oahpa
** try to implement and embed virtual keyboards for each specific Oahpa
** try to learn something about compiling keyboard in general
* dictionaries, generally:
** synchronize the source language entries from a specific dictionary with the
   entries in the morphology component (now especially for sma):
   that means nothing then put the entries from dict that are NOT analyzed
   into the lexC files
** the StarDict on Windows: try the HTML-plugin (that means that users can use
   StarDict 3.2)
** try to reduce the dict-size on mac: experiment with xPointer, etc.
* Fkv:Nob - Nob:Fkv:
** incorporate novel glossary into the dict
** try to implement a web version of the dictionaries in HTML
* KomEngFin:
** test the automatic sorting by Komi alphabet in xsl (as discussed
   with __Trond__)
* SmeNob:
** incorporate the passives into the last version of the sme:nob
** start a new compilation of SmeNob and improve it based on the experience

     gained with SmaNob
* SmaNobSwe:
** extend the smanobswe dictionary: waiting for data (incorporate the data from
   Maja-Lisa)
* SjdRus:
** continue the work at the Kildin-Russian dictionary, next internal deadline
   is after finishing the PhD project description
*** all dict-related work: todo
Ciprian Gerstenberger

* MT
** test A-ITE on Windows
*** todo
Ciprian G…

!!Maja
* prepare: Drøfting og diskusjon om lånord i sørsamisk
**ongoing - with Ove Lorentz - lage oppgaver for gruppearbeid
Maja Lisa Kappfjell
* more work on {{sma}} adjectives
**not prio
Maja Lisa Ka…
* look at incoming loanwords
**not prio
Maja Lisa Ka…
* continue gathering {{sma}} corpus texts
**not prio
Maja Lisa Ka…
* finish compound tags for adjectives
** Ongoing
Sjur Maja Lis…
* Rette goldstandard doc - så mange som mulig - high prio! - min. 1000 (helst
  5000) ord løpende tekst
**ongoing
Maja Lisa K…
* make a list of participants in {{$GTPRIV/}}
**done
Maja L…
* call a select few that have not answered but that we want there
**done
Maja L…
* fix goldstandard files (location)
**ongoing
Maja Lisa K…
** free documents: {{$FREECORP/goldstandard/orig/sma}}
** bound documents: {{$BOUNDCORP/goldstandard/orig/sma}}

* test conversion to xml to ensure there are no markup errors
**
M.
* continue gathering {{sma}} corpus texts
** get {{sma}} articles in Š-bláđđi
** the Gun Utsi book is almost there - one contract missing
***not prio
Maja Lisa Kappfjell
* Corpus gathering:
** contact Inga Margrethe Bjørn Eira
**not prio
Maja Lisa Ka…
**follow up the yearbook-articles
**not prio
Maja Lisa Ka…
** restart the letter mailing thing using WebSak
**not done
Maja Lisa Ka…
* adjectives
* [fix bugs!|http://giellatekno.uit.no/bugzilla]

!!Sjur
* make risten2 budget proposal to the SD

** done
Sjur
* add checkout of the free corpus to the gtsetup.sh script
** not yet
Sjur
* add checkout of the bound corpus when on victorio
** not yet
Sjur
* test the {{sma}} speller on the gold standard document
** done, committed, and sent to Thomas and Maja for checking
Sjur
* add voikko support to our proofing test bench
** not yet
Sjur
* add all our Sámi analysers to voikko and test them as spellers
Sjur

** not yet
Sjur
* run speller tests using Hunspell, Voikko, Polderland for our Sámi lexicons
Sjur
** not yet
Sjur
* install & configure the Unison news reader
** not yet
Sjur
* difftest for fst and PL speller
** not yet
Sjur
* Northern areas
** plan a meeting/seminar in Tromsø
** make a plan for the first two years, update overall plan
* @TTS: continue public tender process
* make __Leif Åge__ send out CD's to distribution points
* contintue Nordplus Sprog project
** more work on the test bench
Sjur
* Write a formal letter to ''Davvi girji'' about electronic dictionaries
* make XSL script to transform leaflet Forrest XDocs to an OOo Drawer document
* name db/risten.no
** budget for risten2, see above
Sjur
* follow-up on some Polderland-related bugs: 621, 630, 652
* find and contact the correct person in SD, to get the manuscript for all Sámi
  teaching material now being given out to the schools
* when the new build infrastructure works as it should, delete the old ones
** done? (should be by Tomi)
Sjur
* [fix bugs!|http://giellatekno.uit.no/bugzilla]

!!Thomas
* prepare: Drøfting og diskusjon om lånord i sørsamisk
** not done
Thom
* test the new build commands and their fst's: {{make NEW-fst TARGET=sme}} etc.
** done
Thom
* check extra hyphens in the lexc continuations with Trond
** not done
Thom
* check for a new computer
** done, it should be here this week
Thom
* fix goldstandard files (location)
** maybe
Thom
* test conversion to xml to ensure there are no markup errors
** just begun
Thom
* prepare text´s about normativity issue to SGL/SGM

** not done
Thom
* Digitalize south saami books
** done
Thom
* [fix bugs!|http://giellatekno.uit.no/bugzilla]
** not done
Thom

!!Tomi
* run tests using Hunspell, Voikko, Polderland for our Sámi lexicons
** not done
Tomi
* put together the TTS preprocessing transducers and scripts
** not done
Tomi
* try to compile voikko
** not working
Tomi
* add all our Sámi analysers and test them as spellers
** not done
Tomi
* when the new build infrastructure works as it should, delete the old ones
** deleting now
Tomi
* document how compounding is controlled in the PLX conversion
** not done
Tomi
* fix double hyphen bugs
* fix PL {{smj}} hyphenator bug
* fix PL conversion bugs
* [fix bugs!|http://giellatekno.uit.no/bugzilla]

!!Trond
* check extra hyphens in the lexc continuations with Thomas
** Done some, but no meeting with Thomas yet
Trond Trosterud
* Course plan for Barents meeting/seminar, october. Financing must be in place.
** Awaiting
Trond Trosterud
* Northern areas
** plan a meeting/seminar in Tromsø
** Awaiting
Trond Trosterud
** make a plan for the first two years, update overall plan
** Awaiting
Trond Trosterud
* MT/Terminology
** Much work here, both on smenob, finsme and smesma
Trond Trosterud
* [fix bugs!|http://giellatekno.uit.no/bugzilla].
** Done 2.
Trond Trosterud

!!!Oahpa!
Status:
Sjur
* sma-oahpa already started, installation will be done by this friday
Ciprian Gerstenberger
* __Ciprian__ should add just 15 entires per pos n, v, a
Ciprian Gerstenberger
* the interim lexicon format as already proposed via email:
Ciprian Gerstenberger
** the current gt_dict lexicon having each relevant meaning group
Ciprian Gerstenberger
   extended with application relevant slots.
Ciprian Gerstenberger

__TODO__
* Register oahpa.no (__Trond__)
* From the start: sjd_oahpa Leksa in deu and eng as well (__Ciprian__)
** delayed
* clock and date for Numra (__Ciprian__)
* activate numra for Skolt Sámi (__Ciprian__)
* email notification when the server goes down (__Ciprian__)
* Finding a volunteer to translate the sme Leksa lexicon to Swedish (__Trond__)
** found, and now returned to do something.
* add Captcha for the feedback e-mail address (__Ciprian__)

!!!Corpus gathering

National Library: there is a meeting in September, to which Trond is going.
Trond will request a couple of Sámi OCR-ed texts (together with the {{nob}}
texts we already have), and use them as test cases for spell-checking and
(automatically) correcting the OCR-ed texts.

This work should be incorporated in the test bench project, and serve as a
basis for developing a better {{nob}} (and potentially {{nno}} as well)
morphologies, to be used in our different projects. The main project,
developing an OCR-adapted Norwegian spell-checker (as open source) should be
financed by the Language bank.

There are starting points both in our own svn and in Apertium, both based on
the lexicons from UiO/Norsk ordbank.

Our main goal with this would be to help the Language bank and the Natinal
Library provide us with usable Sámi (and possibly Norwegian) corpus material,
but there are a number of other secondary goals and positive side effects of
this as well.

__TODO:__
* get all the latest Sámi Parliament protocolls into our repository (__Børre__)
** mostly done, protocolls from the period 2006-2010/1 do not seem to be
Sjur
    translated into {{sme}} - Børre will check this
Sjur
* contact the Sámi translators at the government and ministeries to get more
  parallell texts (__Børre__)
* continue gathering {{sma}} corpus texts (__Maja__)
** get {{sma}} articles in S-bláddi
** the Gun Utsi book is almost there - one contract missing (__Maja__)
* write formal letter to Davvi Girji (__Sjur__)
* send a copy of the signed contracts back to the authors, translators and
  publishers, accompanied by the thank-you letter
* find and contact the correct person in SD, to get the manuscript for all Sámi
  teaching material now being given out to the schools (__Sjur__)
* get the {{sma}} yearbooks from ''Saemien sijhte'' (__Maja__)
* contact certain {{sma}} writers (__Børre__)
* contact Inga Margrethe Bjørn Eira (__Maja__)
* give Maja a WebSak intro (__Børre__)
* restart the letter mailing thing using WebSak (__Maja__)

!!!Promoting Divvun

__Rune Fjelheim__ suggested last week that we should start a separate project
to inform about, teach about and install our language technology tools (not
only Divvun). This would also make a very good feedback channel.

__TODO:__
* make leaflet to inform about the project (__Børre__)
** add InDesign text (__Sjur__)
** make XSL script to transform Forrest XDocs to an OOo Drawer document
    (__Børre, Sjur__)
* distribute CD version through the library bus, the language centres and
  common sami centres in all of Sápmi. Gaaltije in Östersund for example.
  (__Leif Åge, Sjur__)
** make him send out CD's accordingly (__Sjur__)
* update online download log statistics page (__Børre__)

!!!Future plans, directions and ideas

See a separate document in {{plan/strat/5year.jspwiki}}.

!!Northern areas project

First major obstacle: make working keyboards and fonts.
In principle, there are two approaches to keyboards:

Make the Russian-based keyboards work
Design new Kildin-optimized keyboards

The first goal has priority. We should have text analysis in order to consider
whether the second goal is feasible. Trond to follow up test machines.

* Choice of fonts
* Rendering of fonts in MS Word and other important programs
* Keyboard layout
** Empirical phase: What keyboards are around
** Design phase: Make the optimal keyboards -- for all OSes

Write trustworthy and detailed documentation (in Russian)

What we know:
* you can type on some computers (Michael R: it works), not on others. Q: what
  are the differences between the two types of computers?
* some fonts can't display the chars, others can (but usually ugly)

__TODO:__
* Course plan for meeting/seminar, october. Financing must be in place.
** check the time needed to tet visas with the UiT admin (__Trond__)
Sjur

** call the ministry (__Trond__)
Sjur   Trond Trosterud
* Attend a beginners' course in Russian (priority: the alphabet!) near you..
* make a plan for the first two years, update overall plan (__Trond, Sjur__)
* Follow up Windows test machines

!!!Infrastructure

Out of the box experiences:
* fao could not be compiled out of the box (the Oslo gang tried) because of a
  missing {{tmp}} dir
* our list of software to install is incomplete or lacking installation
  instructions
* external tools require manual installation
* there are bugs in the gtsetup.sh script

!!Updated corpus online

See __Ciprian´s__ document about the corpus content in
{{$GTPRIV/plan/corpus/oslo_corpus_update_todo.txt}}.

Issues:
* filenames
* organisation of subdirs?
* content (or lack thereof) of original and converted files
* svn repo reorganisation

```
{{{
facta$convert2xml.pl --nolog --corpdir=/usr/local/share/corp L1all0rt.correct.txt
}}}
```

Error message:
```
{{{
sh: /home/sjur/gtmain/gt/script/text_cat: No such file or directory
L1all0rt.correct.txt: ERROR errors in /home/sjur/gtmain/gt/script/text_cat -q \
  -x -d /home/sjur/gtmain/gt/script/LM "/usr/local/share/corp/tmp/L1all0rt.correct.txt.tmp0":
}}}
```

{{text_cat}} isn't part of our repository, we need to add it - we are using a
modified version of the original file. It should be added with a short README
and some license info, pointing to the original.

__TODO:__

* make the conversion scripts work with the new svn repository
  (__Børre, Ciprian__)
** add check for almost empty content
*** already implemented in convert2xml.pl. Still empty or almost empty files
    are converted. This is because of unexpected structure of content, so that
    the conversion goes wrong, and leaves no real content in the resulting xml
    file.
** add processing of svg files
*** not done
* reorganise subdirs as needed (__Ciprian, Trond__)
** not yet done
* identify parallel nob files - should be automatic, but needs to be checked
  (__Ciprian__)
* sentence-align sme with nob (__Ciprian__)
** the aligner wasn't working, __Børre__ has tried to fix it
*** not yet fixed - bug not found
* analyse/disambiguate (__Ciprian__)
* preferably also dep -> check with Oslo if it can be used (__Ciprian__)
* install in Oslo (__Ciprian, Trond__)
* add better handling of unknown strings in our analysers (__???__)

Corpus infra remake

__Ciprian__ has moved all code relating to {{convert2xml.pl}} to a subfolder.
Still needs testing especially on victorio.

__TODO:__
* convert or move the files in the {{upload}} dir to the real corpus repo
  (__Børre__)
** almost finished, moved files to both {{bound/}} and {{free/}}, will soon
   write documentation on how to use the new corpus infra
* turn on corpus summary crontab (__Børre__)
* needs to check the conversion from old rcs corpus repository to new svn-based
  repository (__Ciprian__)
* add checkout of the free corpus to the gtsetup.sh script (__Børre, Sjur__)
* add checkout of the bound corpus when on victorio (__Børre, Sjur__)
* add updates of free & bound to the svnup alias (__Børre__)

License

__TODO:__
* install & configure the Unison news reader (__Sjur__)

Corpus interface

This depends on the infrastructure cleanup.

__TODO:__
* make a simple web search form for the UiT corpus repository (__Ciprian, X__)
* check out the new version of CWB with Unicode support (__Ciprian__)
** fall

Makefile + tag simplification

Problems with the proofing tools compilation, now solved.

__TODO:__
test the new build commands (__Sjur, Trond, Lene, Thomas__)
kommando: {{make TARGET=sme NEW-fst}} etc.
test by comparing the output of the old and new analysers - it should be
   identical. Test both analysis and generation (ie all your favourite
   usecases)
when the new build infrastructure works as it should, delete the old ones
  (__Sjur, Tomi__)
tests seem fine: Thursday morning August 19, Tomi will implement this.

HFST compilation

Our HFST transducers do still contain all the extra tags resulting from the tag
conversion (cf above). The Apertium versions do not, however. The main problem

here is the forking of smi transducer builing and maintenance - one branch in
Giellatekno, and another in Apertium. This is not good for future maintenance,
and should be corrected ASAP. The tag-removing make steps should be
incorporated in the giellatekno svn.

!!General list

Meänkieli adaptions in our infrastructure.

Requirements:
* separate Subversion repository
* structured roughly as our main repository
* limited access to the closed repo, in the sense that one language group
  should not be able to see the work of another language group (exception for
  administrators) - otherwise it isn't really secret/closed. That is, the FIT
  gang should only see FIT files, and a hypothetical e.g. Estonian gang should
  only see the Estonian files, not the FIT files, etc. But to the extent that
  there are (links to) shared resources the (e.g. links to infrastructure
  things in another place), they should be visible to all groups
* all tools, dtd's, configs work as in and from the main Subversion, such that
  we don't end up with multiple copies. That is, only content files and no
  infrastructure files should be in the closed repo - all the rest should stay
  in the main repo
* we should have scripts that move a full language dir from one repo to the
  other, in both directions, keeping svn history etc.
* the amount of work for setting this up should be minimal, or we have to ask
  them for money

Tentative task list
* figure out the svn dir structure in the closed lang repo, set it up
* figure out how to make the infrastructure in the open/main repo available for
  work in the closed repo (we want to have only one dtd, one XXE config, one
  makefile, etc.)
* when we have an idea of the amount of work, decide whether we just do it, or
  whether we bill them
* implement the previous point
* transfer the language files, preferably using a script for automatic transfer
  and deletion in the open repo
* inform them about the changes

To accommodate future enhancements in different directions (in rough order of
importance):

test bench for all parts of our language technology efforts
test bench enhanced, but not yet complete
improve Forrest i18n support with static sites
reorganise the documentation:
differ between target groups
get better grouping
decide what to write in Forrest and what in wiki
   (cf. [Apertium|http://www.apertium.org/] and
   [http://xixona.dlsi.ua.es/apertium/]) for a similar split)
update/add missing parts
migrate lexc lexicons to XML, splitting the task
Name lexica (the Name project)
Dictionaries (already in XML, task is to integrate them)
At least migrate the lexc open POSes (Komi as a pilot case)
change the look of the documentation web
corpus content moved to Max Planck repositories? Norsk språkbank?
update infrastructure to allow content-restricted spellers for special target
  groups

__TODO:__
* add Apertium to the XServe (__Børre__)
Sjur
* make the new SL Server services functional: (__Børre__)
** group calendars
* set up corpus mirroring on the XServe again (__Børre__)
** on victorio, all files are checked out and converted in
Sjur

   {{/home/apache_corpus/boundcorpus|freecorpus/converted}}
Sjur   Børre G… S Børre Gaup   S.
* finish the restructuring and cleaning of the {{script/}} directory

    __Ciprian__)
* infrastructure remake: (__Børre, Ciprian, Sjur, Tomi, Trond__)
** more modularised make / build infra (prepare for smn, sms, sjd, others)
*** look at omorfi for ideas of how to modularise
** merge gt, kt and st into one
** modularised preprocess and spellrelax
** alternatives to make:
*** [SCons|http://www.scons.org/]
*** [makepp|http://makepp.sourceforge.net/]
*** [Maven|http://maven.apache.org/] (for java-projects)
*** [CMake|http://www.cmake.org/] - __Sjur__ looked at it, and have started to
     experiment
* make a test-all target that runs all tests we have (__Ciprian, Sjur, Trond__)
** delayed until we have restructured the make/build process
* define and document testing routines (__Ciprian, Sjur, Trond__)
** delayed until we have restructured the make/build process

!!!*Linguistics*

!!North Sámi

There are a lot of hyphens found in noun paradigms. They should not be there.
Trond and Thomas will look into them.

__TODO:__
* check extra hyphens in the sme lexc continuations (__Thomas, Trond__)

!!Lule Sámi

__TODO:__
* check extra hyphens in the smj lexc continuations (__Thomas, Trond__)

!!South Sámi

__TODO:__
* read through and comment by Wednesday afternoon (__Sjur__)
* adjectives (__Maja__ with __Thomas, Trond, Sjur__)
** two competing naming conventions of continuation lexicons
*** One naming goes ATTRSUFF-PREDSUFF-STEMTYPE
*** One follows the {{sme}} convention of naming key adjectives
*** There are duplicate lexica
** The comparative issue open here and there
*** The ATTRSUFF-PREDSUFF-STEMTYPE lexica now go to EVENCOMP and
     ODDCOMP only.
* finish compound tags for adjectives (__Maja__)

!!!*Name lexicon/risten.no infrastructure*

__Sjur__ made a budget for risten2 with great help from Ciprian and Trond.
Sjur

__TODO:__
* make budget proposal to the SD (__Ciprian, Sjur__)
* find already approved lists, in paper or electronic form (__term team__)
* convert paper lists to electronic lists (__term team__)
* convert lists to standard XML (__Sjur, Tomi__)
* add prepared lists to risten.no (__Sjur, Tomi__)
* fix i18n bug in risten.no/G5 (so they will work without the proper locale
  request) (__Sjur__)
* fix bugs in lexc2xml; add comments to the log element (__Saara__)
* finish first version of the editing (__Sjur__)
* test editing of the xml files. If ok, then: (__Sjur, Thomas, Trond__)
* make terms-smX.xml <== automatically from propernoun-sme-lex.xml (add
  {{nob}} as well) (the morphological section should be kept intact, in e.g.
  propernoun-sme-morph.txt) (__Sjur, Saara__)
* convert propernoun-($lang)-lex.txt to a derived file from common xml files
  (__Sjur, Tomi, Saara__)
* implement data synchronisation between [risten.no|http://www.risten.no] and
  the cvs repo, and possibly other servers (ie the G5 as an alternative server
  to the public risten.no - it might be faster and better suited than the
  official one; also local installations could be treated the same way)
* start to use the xml file as source file
* clean terms-sme.xml such that all names have the correct tag for their use

(e.g. @type=secondary) (__Thomas, linguists__)
* merge placenames which are errouneously in different entries: e.g. Helsinki,
  Helsingfors, Helsset (__linguists__)
* publish the name lexicon on risten.no (__Sjur__)
* add missing parallel names for placenames (__linguists__)
* add informative links between first names like Niillas and Nils
  (__linguists__)

!!Dictionaries

__Ciprian__ has made a proposal for unifying Oahpa and general dictionary
formats into one DTD.

StarDict is useless for Cyrillic languages, mainly because of the scanning
function (ie point and look-up). We need to find an alternative to StarDict.
Also, the Kildin Sámi users have different needs than we have had in mind. And
they don't have access to a Kildin Sámi keyboard in Windows.

* we need a Russian Windows machine, bought in Russia, using the software they
  use
* we also of course need a western/Norwegian Windows machine for testing

Released:
* SME:NOB update release

Other things dictionary-related:
* risten.no data as part of our dictionaries
* dictionaries as part of risten.no
* dictionaries and risten.no as part of Autshumato ITE

__TODO:__
* kom:fin-eng
** moved the original kom-lex.xml to the inc-dir and froze it
** split it by pos into the working_file dir, the ONLY place to work with
   the dictionary entries
** now, the lexC files are generated via XSLT sheets, no perl scripts
** adjusted the Makefile
** prepared the pipeline for compiling the mac dict
** todo: make a pipeline for StarDict also (as far as I know, __Jaska__ has)
   a Linux machine
* set up risten.no on eXist/XServe (as a beta version site) (__Sjur__)
* set up required infra for smenob on risten.no/XServe (__Sjur__)
* Continue the dictionary infrastructure discussion (__Ciprian, Sjur, Trond__)
* end user documentation (how to download and install) (__Ciprian, Trond__)
* Contact ''Davvi Girji'' about cooperation on electronic dictionaries
  (__Sjur__)
* developing the mobile phone version of smenob:
** [http://pfed.info/wksite/]
* Komi
** take out the doublets to a separate file (__Ciprian__)
** merge the doublets (__Jaska, Trond__)
** Completing the automaton to some state (__Trond, Jaska, Paula__)
* make the sort XSL script available for all languages to keep the source files
  sorted

!!!Proofing tools

Spelling feedback from Malta:
* use FST to model suggestions (Krister et al)
* Lene is evaluating the output of the Polderland speller
** first-letter errors rare, but the PLX speller still changes it quite often
** Oahpa and Divvun approaching each other
* using Wikipedia as a source for spelling errors and test material (a French
  group has made a setup of tools to automatise most of this)

!!South Sámi

Beta release: Aug 25. Contract now ok by Knowledge Concepts. Suggested
timetable sent this morning.

There are no goldstandard files in the correct location in the corpus
repository. This needs to be fixed before we start testing the {{sma}} speller.

__TODO:__
* fix goldstandard files (location) (__Maja, Sjur, Thomas__)
** free documents: {{$FREECORP/goldstandard/orig/sma}}
** bound documents: {{$BOUNDCORP/goldstandard/orig/sma}}
*** done
Sjur

* test conversion to xml to ensure there are no markup errors
  (__Maja, Sjur, Thomas__)
** done for some
Sjur

* test the {{sma}} speller on the gold standard document (__Sjur__)
** done
Sjur

* difftest for fst and PL speller (__Sjur__)
* External beta testers:
** David
** Jovsset
** the Røros group

!!HFST- and Voikko-based proofing tools

__Sjur__ met with the HFST people yesterday. Two things happening in parallell:
hfst3 - being made ready for public release, with proper inclusion of
         autotools, and updated documentation
speller/lookup library:
speller/lookup library
voikko integration of this library

__TODO:__
* Change the license tag to GPL for voikko inclusion. (see above)
** done
* run tests using Hunspell, Voikko, Polderland for our Sámi lexicons
  (__Sjur, Tomi__)
* check out the voikko code, see [this page
  |https://sourceforge.net/apps/trac/voikko/wiki/libvoikko/IntegrationTesting]
  (__Tomi__)
* try to compile it (development is done on Linux, no MacOS X testing so far)
  (__Tomi__)
* add voikko support to our proofing test bench (__Sjur__)
* add all our Sámi analysers and test them as spellers (__Tomi, Sjur__)

!!Speller bugs

List of bugs returned from Polderland:
* 621
* 630
* 652
* 656
* 676

Tag reordering for abbreviations have caused a lot of problems:

{{{
smj:
hr.
hr. hr+ABBR+Acc
cand.philol.
cand.philol.   cand.philol+ABBR+N+Acc
Per
Per Per+N+Prop+Mal+Sg+Attr

sme:
hr.
hr. hr+N+ABBR+Acc
Per
Per Per+N+Prop+Mal+Sg+Attr
}}}

Open issues based on test results:

!sme
Version: __Davvisámi, version 1.2, 2009-09-18__

* 399 - missing numerals (plural forms) - still __OPEN__
* 425 - X not recognised; single letters were left out - still __OPEN__
* 435 - roman numbers - inflection of single letter numbers
        rejected, as well as some complex numbers (but is ok in {{smj}}) -
        still __OPEN__
** we should pregenerate all numbers once and for all, and store them in a
        separate lexicon file
* 461 - __REGRESSION:__ missing suggestion (sáhkki)
* 508 - __REGRESSION:__ accepts smj entries (most likely abbreviation missing)
* 520 - __REGRESSION:__ r9 and š9 not defined (abbr. missing)
* 595 - prefix+name without hyphen (''ovdaLot'' instead of ''ovda-Lot'') -
        still __OPEN__
* 603 - suomabealdi accepted - still __OPEN__
* 606 - compound-tags LEXICON VUOHTA - still __OPEN__
* 613 - short gen. as second compound part - still __OPEN__
* 619 - numerals and pronouns to NAMÁK and SASJ fails - ''vihttasoarttat''
        remaining - still __OPEN__
* 629 - ''a'' taking part in compounding without hyphen - still __OPEN__
** only open case has word ''A-finálaid'' compounded
* 647 - numerals+NOUN - still __OPEN__, open case has uppercase letters
* 648 - unmotivated suggestions with numeral+noun - still __OPEN__
* 661 - __REGRESSION:__ abbr. not recognized
* 709 - sámedikkeválga accepted - __OPEN__
* 728 - vowel shortening GenCmp+Left-tagged - still __OPEN__
* 779 - caseforms of pronoun okatahat - still __OPEN__
* 785 - does not recognize alphabet-abbr+noun - __OPEN__
* 802 - __NEW:__ multiword propernouns
* 803 - __NEW:__ FINJU- words accepted single-handed
* 804 - __NEW:__ guovttilogát, njealjilogát
* 805 - __NEW:__ Nouns+acronyms

!smj
Version: __Julevsáme, version 1.2, 2009-09-20__
* 435 - roman number - single letter numbers now recognised
**   we should pre-generate all numbers once and for all, and store them in
        a separate lexicon file
**   please note that ''inflection'' of single letter numerals is __fine__
        in {{smj}}, as opposed to {{sme}}
* 482 - ''polardutkamin'' not recognized - __FIXED__
* 496 - __REGRESSION:__ unrecognised clitics
* 556 - non-existent word accepted - __FIXED__
* 594 - ''lågenanguoktáj'' not recognized - still __OPEN__
* 595 - __REGRESSION:__ prefix+name as split comp without hyphen
* 596 - C-giellan is not accepted - still __OPEN__
* 600 - Gen+hyph compound - __FIXED__
* 627 - prefix + hyhpen does not get accepted - __FIXED__
* 647 - numerals+NOUN - still __OPEN__, open case has uppercase letters
* 648 - unmotivated suggestions with numeral+noun - still __OPEN__
* 650 - __REGRESSION:__ noun prefix+name compound without hyphen
* 652 - UPPERCASE-typos only get acronym-suggestions - still __OPEN__
* 692 - numeral-variants - all but one fixed (''gáktsalågenantjuotakta''), but
        still __OPEN__
* 744 - __REGRESSION:__ numerals + clitic
* 803 - __NEW:__ VINJU- words accepted single-handed
* 805 - __NEW:__ Nouns+acronyms

__TODO:__
* document how compounding is controlled in the PLX conversion (__Tomi__)

!!Hyphenator bugs

Open issues based on test results :

!sme
Lexicon version: __Davvisámi, version 1.2, 2009-09-18__

No known issues!

!smj
Lexicon version: __Julevsáme, version 1.2, 2009-09-20__
* 670 - Hard hyphen replaced with soft hyphen: ''10-biejvvásattja'' (the word
        is not rec.; Bug #711) - still __OPEN__

!*sma*

Command to test the hyphenator:

```
{{{
preprocess dev/corp/pressemelding.txt | lookup bin/hyph-sma.fst | cut -f2 | \
lookup bin/hyphrules-sma.fst | grep -v '^$' | cut -f2 | uniq | see
}}}
```

__TODO:__
* fix PL hyphenator errors (__Tomi__)
** almost done - one {{smj}} bug left

Installer changes

__TODO:__
* test InDesign installer (__Sjur__)

User documentation

__TODO:__
* InDesign documentation (__Sjur__)
** Norwegian translation received from Davvi Girji

1.2 release

Content:
* several smj bug fixes
* lexicalisations
* InDesign Mac & Win
* new OOo beta
* improved installers, at least for Mac, preferably also for Windows

2.0 release

Content:
* official {{sma}} proofing tools release

Nordplus Språk proofing test bench project

Arno Teigseth is hired for the summer. He is sitting in Equador, 7 hours after
Norway. He is working with a quechua speller (using hunspell) and dictionary.

A goal is to use texts from parallel domains, in order to be
able to compare nob and sme. We start out with 100000 words, 5000
for each genre:
Discussion fora on the net
Blogs

Discussion fora can be problematic, as they contain a lot of very oral language
with deliberate oral or dialectal forms used in the text. As such, the texts
won't be typical for the kind of texts people would like to spell-check, and
thus the genre is problematic for the purpose at hand.

Possible other genres:
Minutes from local associations
Pupil texts - look for the home pages of the schools
exam texts by 10th grade pupils, where the pupils have published the texts
  themselves
skrivebua - Nordland fylkeskommune: all three Sámi languages

Text to speech

There will be regular meetings in this project from now on, every second week.

__TODO:__
* put together the preprocessing transducers and scripts (__Tomi__)
* refine syntax / dependency rules (__Biret Ánne__)
* continue public tender process (__Sjur__)

Machine Translation

Background:

Kevin and Ryan have now closed the work on smenob and finsme. They will
continue, but mainly do other things. As for smesmj, the possible student
input is rescheduled for October.

We hope to have a student working one month on smesmj this summer.
The world now knows (vaguely) that we work on MT, so we should do, as well.
We also need to work together with the external ones.

All languages

__TODO__

* Set up hfst and apertium for all relevant machines
* Improve the sme-dis for MT purposes
* Go through the bilingual transfer lexica for more general translations
* Add alternative translations and selection rules
* Arrange an Apertium seminar (later)
* transfer hfst build infra from Apertium to Giellatekno

finsme

Technical problems with compunds in Finnish: each part in the compound is
separated by a space character (+Ux20), which is creating problems in the
transfer.

__TODO__
* Tag unification
* Adp split in fin (__Trond__)

smenob

Needed: Lexicon completion. Grammatical transfer rules

smesmj

New worker underway.

smesma

Infrastructure in place, to demo in Trondheim to encourage more work and
projects.

CAT

__Ciprian__ has made a translation memory based on our parallel corpus. The
test looks promising, he has identified certain issues. We need to be clear
about the license of the original parallel texts, and we need to know the
original language (ie translation direction).

__TODO:__
* make intelligent nobsme risten.no - requires {{nob}} morphology
** nob.fst a la Trond
** norskordbank.fst
** nob.fst based on an xml source synchronised with Ordbanken source texts
** should be coordinated with the Apertium analyser
* write a TM script to turn a parallell corpus into TM entries
** testing A-ITE on Norwegian Windows

SMA seminar in August/September

* Place: Trondheim
* Date: 30.8.-3.9.

Theme:
* editing dictionaries
* beta tools for {{sma}}
* orthographic questions

__TODO:__
* book lunch, coffee
* make a list of participants (12 + 5 + 3) in {{$GTPRIV/}} (__Maja__)

Trond Trosterud

** done
Sjur

* call a select few that have not answered but that we want there (__Maja__)
** done
Sjur

* preparations:
** test installation of all needed software (__Ciprian, Børre__)
** test XMLEditor on Windows, including plugin installation
   (__Ciprian, Børre__)
*** plugin/configuration not yet tested, will need an installation batch/shell
    script for Windows
** try to work out a unified DTD for dictionaries and Oahpa (or augment the
   general dictionary DTD with Oahpa-required features) (__Ciprian__)
** sma Divvun beta - see above

!!!*Other*

!!Start to make yearly reports

__TODO__
* Set up a catalogue in plan/ (__Trond__)

!!Thursday inhouse seminar

Next time suggestion list:
introduction to xslt - __Ciprian__ to start out
relevant xslt issues:
basic principles of xslt …
sorting in xslt … (have a look at the dictionary sort xslt script)
converting from one xml format to another wilt xslt (sugg: convert from
    DivvunGT dictionary dtd to MacDict xml)

Future seminars:
XQuery
More XML (needs concretisation)
UML
other suggestions?

!!Summer planning

Topics:

* speller test project -> Sjur, Børre, Thomas, X
* speech synthesis -> Sjur, Trond, BA as a starter
* risten.no -> Ciprian, Sjur, Tomi
* Barents project follow-up meeting -> Trond, Sjur
* sme and smj proofing tools, next version -> Thomas, Tomi, Sjur
* HFST-based proofing tools -> Sjur
* MT terminology project -> Trond, Linda, Fran, Kevin
* HSL centre status -> Trond

Dates:

* August: IceTAL, Reykjavik - Linda, Francis
* {{sma}} seminar in Trondheim: Lexicography, Oahpa, Divvun2 beta
* September 8: Konvens-workshop Saarbrücken: LT and text-technological methods
  for e-learning http://konvens2010.de/workshop_en.html:
** Lene will try to get an abstract accepted, acceptance 6.7

!!!*Next meeting, closing*

The next meeting is 6.9.2010, 09:30 Norwegian time.

The meeting was closed at 10:47.
Sǯ S.

!!!*Appendix - task lists for the next week*

!!Boerre
[iCal|/doc/admin/weekly/2010/Tasks_2010-08-23_Boerre.ics]
* get all the latest Sámi Parliament protocolls into our repository
* contact the Sámi translators at the government and ministeries to get more
  parallell texts

* corpus access on the XServe
* turn on corpus summary crontab
* implement language switch for static divvun site
* improve XSL script to transform leaflet Forrest XDocs to an OOo Draw document
* get translations of thank-you letter
* make the new SL Server services functional:
** group calendars
* set up corpus mirroring on the XServe again
* give Maja a WEbSak intro
* add checkout of the free corpus to the gtsetup.sh script
* add checkout of the bound corpus when on victorio
* {{sma}} seminar in August/September:
** test installation of all needed software
** test XMLEditor on Windows, including plugin installation
** test svn checkout and patch creation on Windows, using a GUI client
* [fix bugs!|http://giellatekno.uit.no/bugzilla]

!!Ciprian
[iCal|/doc/admin/weekly/2010/Tasks_2010-08-23_Ciprian.ics]
* {{sma}} seminar in August/September:
** prepare slides
** test XMLEditor plugin installation
** augment the dictionary DTD with oahpa features and requirements
** finnish the setup of the sma-oahpa on victorio
** build a preliminary data set for sma-Leksa (based on the DTD)
* Skolt-Oahpa:
** fine tuning of semantic tags
** finish localization
* terminology:
** merged the 2 doc files with the xls file containing the 2006 law terminology
   db
** prepare a StarDict test version of the merged data for demo
** compile a dictionary in StarDict format with all data featuring the
   properties wished by Ellen Ravna
* add license header to all files
* PhD-Projektbeskrivelse (internal deadline end of June 2010):
** search, read, write (execute): high priority!
* read and answer the newsgroups messages
* update corpora (both monolingual and parallel) for Oslo (Glossa)
** check the correctness of parallelity between sme and nob files
** sentence-align sme with nob: fix tca2 problem
** analyse/disambiguate: todo -- waiting for the last version of FSTs
* corpus infra:
** check the conversion from old rcs corpus repository to new svn-based
** make the conversion scripts work with the new svn repository
*** add check for almost empty content
*** add processing of svg files
* infrastructure
** test cwb
** continue with restructuring and cleaning the script catalogue as suggested
   in the newsgroup
** transform sme-lexC files into XML format
** make a schema/dtd description of the lexC-file (experiment with
   Komi and Romanian)
* GT web:
** add a tree visualizer for the dependency trees
** add input help for special characters on the tool sites
** automatise the web statistics
** filter (English, German, etc.) input using language detection tools
** put a note on the sites that these are NOT MT tools
** input help for generating wordforms (dropdown menus).
* Sandbox Oahpa:
** integrate reCAPTCHA into Django (as __Lene__ suggested)
** debug the installed sb_oahpa
** update and correct the Oapha docu site
* Running Oahpa:
** implement a testbench for Vasta (as __Lene__ needs)
** integrate reCAPTCHA after the SB-test
** add an Oahpa clock and date excercise (cf. Numra)
** email notification when the server goes down
** check the XXX?
* Sjd/Kom/Sms/Etc-Oahpa
** try to implement and embed virtual keyboards for each specific Oahpa

** try to learn something about compiling keyboard in general
* dictionaries, generally:
** synchronize the source language entries from a specific dictionary with the
   entries in the morphology component (now especially for sma):
   that means nothing then put the entries from dict that are NOT analyzed
   into the lexC files
** the StarDict on Windows: try the HTML-plugin (that means that users can use
   StarDict 3.2)
** try to reduce the dict-size on mac: experiment with xPointer, etc.
* Fkv:Nob - Nob:Fkv:
** incorporate novel glossary into the dict
** try to implement a web version of the dictionaries in HTML
* KomEngFin:
** test the automatic sorting by Komi alphabet in xsl (as discussed
   with __Trond__)
* SmeNob:
** incorporate the passives into the last version of the sme:nob
** start a new compilation of SmeNob and improve it based on the experience
   gained with SmaNob
* SmaNobSwe:
** extend the smanobswe dictionary: waiting for data (incorporate the data from
   Maja-Lisa)
* SjdRus:
** continue the work at the Kildin-Russian dictionary, next internal deadline
   is after finishing the PhD project description
* MT
** test A-ITE on Windows

!!Maja
[iCal|/doc/admin/weekly/2010/Tasks_2010-08-23_Maja.ics]
* prepare: Drøfting og diskusjon om lånord i sørsamisk
* more work on {{sma}} adjectives
* look at incoming loanwords
* continue gathering {{sma}} corpus texts
* finish compound tags for adjectives
* Rette goldstandard doc - så mange som mulig - high prio! - min. 1000 (helst
  5000) ord løpende tekst
* make a list of participants in {{$GTPRIV/}}
* call a select few that have not answered but that we want there
* fix goldstandard files (location)
** free documents: {{$FREECORP/goldstandard/orig/sma}}
** bound documents: {{$BOUNDCORP/goldstandard/orig/sma}}
* test conversion to xml to ensure there are no markup errors
* continue gathering {{sma}} corpus texts
** get {{sma}} articles in Š-bláđđi
** the Gun Utsi book is almost there - one contract missing
* Corpus gathering:
** contact Inga Margrethe Bjørn Eira
**follow up the yearbook-articles
** restart the letter mailing thing using WebSak
* adjectives
* [fix bugs!|http://giellatekno.uit.no/bugzilla]

!!Sjur
[iCal|/doc/admin/weekly/2010/Tasks_2010-08-23_Sjur.ics]
* make risten2 budget proposal to the SD
* add checkout of the free corpus to the gtsetup.sh script
* add checkout of the bound corpus when on victorio
* test the {{sma}} speller on the gold standard document
* add voikko support to our proofing test bench
* add all our Sámi analysers and test them as spellers
* run tests using Hunspell, Voikko, Polderland for our Sámi lexicons
* install & configure the Unison news reader
* difftest for fst and PL speller
* Northern areas
** plan a meeting/seminar in Tromsø
** make a plan for the first two years, update overall plan
* @TTS: continue public tender process
* make __Leif Åge__ send out CD's to distribution points
* contintue Nordplus Sprog project
* Write a formal letter to ''Davvi girji'' about electronic dictionaries
* make XSL script to transform leaflet Forrest XDocs to an OOo Drawer document

* name db/risten.no
* follow-up on some Polderland-related bugs: 621, 630, 652
* find and contact the correct person in SD, to get the manuscript for all Sámi
  teaching material now being given out to the schools
* when the new build infrastructure works as it should, delete the old ones
* [fix bugs!|http://giellatekno.uit.no/bugzilla]

Thomas
* prepare: Drøfting og diskusjon om lånord i sørsamisk
* check extra hyphens in the lexc continuations with Trond
* fix goldstandard files (location)
* test conversion to xml to ensure there are no markup errors
* prepare text´s about normativity issue to SGL/SGM
* Digitalize south saami books
* [fix bugs!|http://giellatekno.uit.no/bugzilla]

Tomi
* run tests using Hunspell, Voikko, Polderland for our Sámi lexicons
* put together the TTS preprocessing transducers and scripts
* try to compile voikko
* add all our Sámi analysers and test them as spellers
* when the new build infrastructure works as it should, delete the old ones
* document how compounding is controlled in the PLX conversion
* fix double hyphen bugs
* fix PL {{smj}} hyphenator bug
* fix PL conversion bugs
* [fix bugs!|http://giellatekno.uit.no/bugzilla]

Trond
[iCal|/doc/admin/weekly/2010/Tasks_2010-08-23_Trond.ics]
* check extra hyphens in the lexc continuations with Thomas
* Course plan for Barents meeting/seminar, october. Financing must be in place.
* Northern areas
** plan a meeting/seminar in Tromsø
** make a plan for the first two years, update overall plan
* MT/Terminology
* [fix bugs!|http://giellatekno.uit.no/bugzilla].